# Machine Learning for Sample Selection Models

Owen McDevitt[*]

May 9, 2019

## Abstract

The existing literature on sample selection models, and problems thereof, is extensive and expanding. However, little has been discussed regarding the use of machine learning as a tool to correct selection bias. Despite this lack of literature, the predictive capabilities of many machine learning algorithms make them an ideal candidate for the task. This paper offers both a primer on the topic and a demonstration of its potential use. To explore the topic, I employ simulated data sets: (1) a large, unbiased, normally distributed reference set as the population, and (2) a biased subset of the population that is ”observed”. I correct for selection bias in the observed data using each observation's predicted probability of selection as determined by four different classifiers: neural net, random forest, naive bayes, and logit regression. I then compare the effectiveness of each classifier in recovering the true value of $\beta_1$ in the corrected model. Though the neural net is most computationally expensive, it ultimately provides the best correction - slightly better than the standard approach of using a logit model. Overall, the results are promising, and they affirm the potential that machine learning has in application to

[*] Department of Economics, University of Oklahoma. E-mail address: omcdevitt1022@ou.edu

sample selection models. However, more extensive simulation studies are necessary to make any definitive conclusions regarding its utility in practice.

# 1 Introduction

Beginning with the seminal work of James Heckman Heckman (1979), economists have produced an ever-growing literature on sample selection bias. Sample selection bias occurs when studying a population of data using a sample that is not wholly representative, that is, a non-random sub sample. The literature provides strategies that combat this bias under a variety of assumptions and in a variety of contexts. However the use of machine learning as a tool for correcting selection bias has yet to be seriously explored. Because machine learning models are often exceptional predictors, they provide an ideal candidate for the task. By exploiting this predictive accuracy, one can obtain the probability that a given observation will be ”selected” or not. The conditional probability of selection dependent on a vector of observed covariates is known as the propensity score Rosenbaum and Rubin (1983). This predicted probability can then be included in a regression with the observed data, and this corrected regression model ultimately addresses the bias. I specifically examine the use of four different classifiers in obtaining the predicted probabilities: a Neural Net, a Random Forest, a Naive Bayes classifier, and a logistic regression. The standard approach uses a logistic regression, so I include it for comparison's sake.

To test the effectiveness of the classifiers, I use simulated data based on a sample selection model. First, I simulate a data set from an outcome equation to be used as the population, and then I subset this population according to a selection equation. This subset contains data that is ”observed”. Due to the non-random sub-setting of the population, the observed data is systematically biased. Thus a linear model using this data produces an incorrect coefficient on the regressor of interest. Therefore, using this model to study the population as a whole will lead to erroneous conclusions. The dangers of failing to make this correction are evidenced by the following canonical example: studying wages.

When studying potential wages of an entire population, selection bias is nearly impossible to avoid. For example, if one has a data set that only contains information for those who are employed, then any study that does not correct for the selection bias will only be able to draw conclusions about wages conditional on employment - not wages in general. Assume that we are studying the effect of education on wages. Also, assume that a given worker chooses to work if his/her wages are above a certain level (reservation wage). In this case, high-education individuals will be well represented because the wage offer they receive will likely be sufficient. However, low education individuals are more likely to be offered wages below their reservation wage. Thus, the only low education individuals that appear in the data set are the ones who are earning sufficiently high wages. Therefore, education is dependent on the error term, causing the effect of education on wages to be biased downward due to a proportion of low education individuals self-selecting out of the workforce. That is, selection into the workforce for low education workers is dependent on the wage offer.

Keeping this in mind, one can see that sufficiently addressing selection bias is vital to making reliable conclusions. The difference between observing wages and observing all potential wage offers is subtle yet paramount. Fortunately, there are many ways to correct for this bias. In this paper, I use classifiers to obtain the predicted probability that a given observation is selected. I then incorporate these probabilities of being selected, or propensity scores, into the model experiencing selection bias. By controlling for the variables that determine selection, I essentially remove the effect of selection from the error term. Thus, it allows us to draw robust conclusions with regard to the entire population and therefore capture the ”true” coefficient on the regressor of interest. That is, the value that I assigned for $\beta_1$ in the outcome equation. This whole process is shown in greater detail in the methods section.

Because selection bias is so pervasive, having a flexible, effective, and easy-to-implement method for correction is essential. Machine Learning as a tool for this correction provides all three of the aforementioned qualities. Most importantly, it is comparatively stronger than most existing methods with regard to flexibility, since it does not necessarily rely on as many assumptions.

# 2 Literature Review

The most well-know method for correcting selection bias is the Heckman correction. By viewing the selection bias as an omitted variable problem, one can control for the dependence between the regressor of interest and the error term using a two-step control function Heckman (1979). The ”omitted variable” in this case is $(\varepsilon_i \mid \eta_i > -\gamma_1 Z_i)$. Essentially, this is the error conditional on whether or not someone is selected. By incorporating this back into the model using the Inverse-Mills ratio, one is able to correct for the bias.

However, Heckman’s solution relies on distributional assumptions for the error terms, specifically that they are jointly normal. When these assumptions are not met, the solution is inconsistent and misleading Goldberger (1983). Additionally, the correction is inefficient when their exists a correlation between the error term and selection mechanism Puhani (2000). Since then, the literature has expanded to weaken the assumptions of Heckman’s original solution and improve its applicability and performance.

Another popular method in correcting for selecting bias is Propensity Score mathcing (PSM) Rosenbaum and Rubin (1983). PSM uses the observed covariates for each observation to estimate the probability of being treated by matching non-treated observations with treated ones. However, PSM implicitly assumes that all factors impacting this selection and the regressor of interest have already been observed. Moreover, it relies on data being available for the non-treated individuals.

The most popular method for estimating propensity scores is logit regression Austin (2011). This paper also estimates the predicted probability of being treated with a logit regression, however it is primarily as a comparative baseline for the machine learning methods. Despite logistic regression being standard, the use of machine learning to estimate propensity scores has been shown to be effective. In a simulation study using a Neural Net compared to a logistic regression in a propensity score matching model, the Neural Net was found to produce less numerically biased estimates Setoguchi et al. (2008). However, up until now, no extensive simulation study has been done on the topic as a whole.

On a broader scale, machine learning is increasingly contributing to economic literature, especially where it relates to econometrics and causal inference Athey (2018). Though the aim of machine learning is fundamentally different than that of traditional econometrics, its predictive strength alone makes it useful in a variety of contexts. Machine learning has been applied to many areas including regression discontinuity, difference-in-differences, structural modeling or individual and firm behavior, and more. Lastly, recent research by Chernozhukov has been particularly instrumental in marrying the two fields - causal inference and machine learning Newey et al. (2017).

Following the lead of similar simulation studies, this paper begins to overview the application of machine learning as it pertains to sample selection models. Eventually, it may serve as a guide to empirical use. This will hopefully contribute to the machine learning/causal inference literature. As of now however, this paper only examines a single context and thus fails to offer a complete view of the topic.

# 3 Data

In order to test the performance of the different classifiers, I simulate a sample selection model. This model includes two simulated data sets: first, a large data set (N = 100000) of unbiased data following a normal distribution, and then a smaller, biased subset of that data chosen according to a selection equation. The larger data set represents the population, and the smaller data set represents the individuals which are observed. On average, half of the total population is observed. The following equations are used for the simulation. Equation (1) is the outcome equation, and equation (2) is the selection equation. Equation (3) uses the selection equation to determine whether a given individual is observed.

The simulated sample selection model can be formalized in the following way:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1}$$

$$U_i = \gamma_0 + \gamma_1 Z_i + \eta_i \tag{2}$$

$$d_i = \{ \begin{array}{ll} 0 & if\, U_i \leq 0 \\ 1 & otherwise \end{array} \tag{3}$$

For simplicity, $\beta_0$ and $\gamma_0$ are assigned values of zero. Thus, there is no intercept on the models. Also, I assign a value of -1 to $\beta_i$. This is the coefficient on our regressor of interest. When I correct for the selection bias, I am trying to obtain this value. Additionally, $X_i$ and $Z_i$ each follow a normal distribution such that $X_i$,$Z_i \sim \mathcal{N}(0,\sigma^2)$ . Also, $\eta$ and $\varepsilon$ follow a bivariate normal distribution, such that $\begin{pmatrix} \varepsilon \\ \eta \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right]$

First, I simulate the population data set of 100000 observations according to the outcome and selection equations. Then, I add a column of treatment values, such that observations are assigned a value of $d = 0$ when $U_i \leq 0$ and a value of $d = 1$ otherwise. Lastly, I subset this data, removing observations where $d = 0$. This results in a biased data set with regards to the relationship between $Y_i$ and $X_i$. The model with only observed data is as follows:

$$ObsY_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{4}$$

As evidenced by Table A1 in the appendix, this model results in a biased value of $\beta_1$. The linear model with observed data (4) has a $\beta_1 = -.812$ and the linear model with population data (1) has a $\beta_1 = -1.001$. Summary statistics for the two data sets are also available in the appendix.

# 4 Methods

In our simulated model, $U_i$ measures the tendency to be selected and $Y_i$ measures the outcome we want to study. Thus, we observe the outcome of any given data point only if the selection variable is positive. We ultimately want to study the relationship between $Y_i$ and $X_i$. However, we cannot yet directly observe this. What we do observe is $E[Y|X_i = X, Z_i = Z, d = 1]$, or, our outcome variable conditional on selection. This can be rewritten as $\beta_1 X_i + E[\varepsilon|\eta > -\gamma Z_i]$. Thus, the expectation of our outcome consists of two factors: the regressor of interest and the expected error caused by selection. Therefore, by controlling for the error caused by selection, we are able to achieve an accurate measure of $\beta_1$.

In order to control for this error, I obtain the predicted probability of being selected for each observation in the data, that is, $P(d = 1 | X_i = X, Z_i = Z)$. By incorporating this predicted probability in the model of observed data, the effect of selection is essentially removed from the error term.

When incorporating these probabilities, I use the utility-maximizing probability. That is, given $d = 1$, I use $p = P(d = 1)$, and given $d = 0$, I use $p = 1 - P(d = 1)$. Additionally, to capture more than just a linear relationship, I incorporate the predicted probabilities as a flexible function. I include $p$, $p^2$, and $p^3$. Thus, our corrected model stands as follows:

$$ObsY_i = \beta_0 + \beta_1 X_i + \beta_2 p_i + \beta_3 p_i^2 + \beta_4 p_i^3 + \varepsilon_i \tag{5}$$

As mentioned, I obtain the predicted probabilities using four different classifiers. I will briefly overview each of the three machine learning methods that I use.

The random forest classifier essentially averages the predictions of many decision trees. Decision trees attempt to minimize error by splitting the data into increasingly small sub-samples. When no tree depth is specified (as is the case with our model), the data is continually split until the sub samples are maximally pure, that is, all observations in the sub-sample fall into a single category. This often results in overfitting. However, the random forest corrects for this by averaging the predictions over many trees - each tree run on a redrawn sample of the population.

The next classifier I use is a nueral network. The nueral net begins with inputs ($X_i$ and $Z_i$ in our case) and, by continually adjusting how the model is weighted, reaches the output ($d_i$). It achieves this by passing through ”layers” that perform transformations on the inputs, ultimately creating some non-linear function mapping the inputs to the output. In addition, I perform cross validation on my neural net.

Lastly, I use a naive bayes classifier, which simply uses Bayes theorem to calculate the posterior probability of each potential outcome ($d_i$) and then chooses the outcome with the greatest probability. The naive bayes is simple and thus computationally inexpensive.

In order to compare the effectiveness of the classifiers, I simply look at which ones result in a model with a $\beta_1$ closest to the true value when incorporated in the corrected model. I also run the classifiers several times and compare their computational efficiency.

# 5 Results

When running the corrected models, all four classifiers are able to recover a value of $\beta_1$ that is within .1 of the true value. However, the neural net and logistic regression performed generally better than the other two classifiers, with the neural net performing best. The results of the corrected regressions using machine learning are contained in Table 2 of the appendix. (1) is the Naive Bayes, (2) is the nueral net, and (3) is the Random Forest. Also, contained in table 3 of the appendix are the results from the logistic regression correction.

I ran each of the four algorithms 100 times. A violin plot containing the computation times for each one is shown in the appendix as well. These plots show the probability density of computation times for each method. The neural net was the most computationally expensive, and the random forest was close behind. The logistic regression and naive bayes did not come with a significant computation cost.

The results tentatively confirm that machine learning - especially neural nets - offer an effective alternative in correcting for selection bias. Moreover, the machine learning algorithms rely on relatively few assumptions. Therefore their application is not as constrained, and it offers a more flexible solution to selection bias relative to existing methods. However, because the standard approach of using logistic regression is only marginally outperformed by the much more computationally expensive neural net, there may not be immediate justification for its use. This paper is limited by the fact that corrections were only tested in a single context. Thus, it is difficult to generalize the findings.

## 6 Conclusion

This research offers a preliminary look at the use of machine learning in sample selection models. Though the results are promising, more research is needed to make definitive conclusions. However, this paper tentatively concludes that using a nueral net classifier results in the best correction. Moreover, the use of a random forest or naive bayes classifier removes a portion of the selection bias but is not nearly as effective.

Future research can increase the number of classifiers, the number of scenarios each classifier is tested under (what assumptions are made), and the number of simulations being run. Some more potential thoughts to examine in closer detail are: an IV or many IVs being included in the model, varying dimensionality of X, varying strength of the selection bias, and varying portion of the population being selected.

Looking at how machine learning classifiers perform under the aforementioned conditions will allow us to gain a clearer idea of where it belongs in application. However, I do believe it will ultimately prove to be empirically useful.

# References

Athey, Susan. 2018. "The Impact of Machine Learning on Economics." .

Austin, Peter C. 2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioral Research* 46 (3):399424.

Goldberger, Arthur S. 1983. "Abnormal Selection Bias." *Studies in Econometrics, Time Series, and Multivariate Statistics* :6784.

Heckman, James. 1979. "Sample Selection Bias As a Specification Error." *Econometrica* 47 (1).

Newey, Whitney K., Christian Hansen, Esther Duflo, James Robins, Victor Chernozhukov, Mert Demirer, and Denis Chetverikov. 2017. "Double/debiased machine learning for treatment and structural parameters." .

Puhani, Patrick. 2000. "The Heckman Correction for Sample Selection and Its Critique." *Journal of Economic Surveys* 14 (1):5368.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1):4155.

Setoguchi, Soko, Sebastian Schneeweiss, M. Alan Brookhart, Robert J. Glynn, and E. Francis Cook. 2008. "Evaluating uses of data mining techniques in propensity score estimation: a simulation study." *Pharmacoepidemiology and Drug Safety* 17 (6):546555.

# A Mathematical Appendix

Table 1:

| | *Dependent variable:* | |
|---|---|---|
| | Y | obsY |
| | (1) | (2) |
| X | $-1.001^{***}$ | $-0.812^{***}$ |
| | (0.002) | (0.004) |
| Constant | $-0.002$ | $0.319^{***}$ |
| | (0.002) | (0.004) |
| Observations | 100,000 | 49,755 |
| $\mathrm{R}^2$ | 0.665 | 0.493 |
| Adjusted $\mathrm{R}^2$ | 0.665 | 0.493 |
| Residual Std. Error | 0.709 (df = 99998) | 0.661 (df = 49753) |
| F Statistic | $198{,}902.500^{***}$ (df = 1; 99998) | $48{,}327.610^{***}$ (df = 1; 49753) |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 2:

| | *Dependent variable:* | | |
|---|---|---|---|
| | obsY | | |
| | (1) | (2) | (3) |
| X | $-1.084^{***}$ | $-1.002^{***}$ | $-0.958^{***}$ |
| | (0.005) | (0.004) | (0.004) |
| poly(pred.1st.best, 3)1 | $-100.676^{***}$ | $-91.238^{***}$ | $-77.010^{***}$ |
| | (1.282) | (1.031) | (1.010) |
| poly(pred.1st.best, 3)2 | $23.641^{***}$ | $9.305^{***}$ | $4.745^{***}$ |
| | (0.892) | (0.884) | (0.898) |
| poly(pred.1st.best, 3)3 | $-3.153^{***}$ | $-8.495^{***}$ | $-8.300^{***}$ |
| | (0.886) | (0.881) | (0.892) |
| Constant | $0.161^{***}$ | $0.210^{***}$ | $0.236^{***}$ |
| | (0.004) | (0.004) | (0.004) |
| Observations | 49,930 | 49,930 | 49,930 |
| R$^2$ | 0.569 | 0.571 | 0.553 |
| Adjusted R$^2$ | 0.569 | 0.571 | 0.553 |
| Residual Std. Error (df = 49925) | 0.610 | 0.609 | 0.622 |
| F Statistic (df = 4; 49925) | $16{,}458.040^{***}$ | $16{,}585.240^{***}$ | $15{,}415.450^{***}$ |
| *Note:* | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 3:

| | *Dependent variable:* |
|---|---|
| | obsY |
| X | $-1.004^{***}$ |
| | (0.004) |
| poly(pred.1st.best, 3)1 | $-91.509^{***}$ |
| | (1.029) |
| poly(pred.1st.best, 3)2 | $6.666^{***}$ |
| | (0.888) |
| poly(pred.1st.best, 3)3 | $-10.254^{***}$ |
| | (0.885) |
| Constant | $0.209^{***}$ |
| | (0.004) |
| Observations | 49,930 |
| $R^2$ | 0.570 |
| Adjusted $R^2$ | 0.570 |
| Residual Std. Error | 0.609 (df = 49925) |
| F Statistic | $16{,}577.450^{***}$ (df = 4; 49925) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

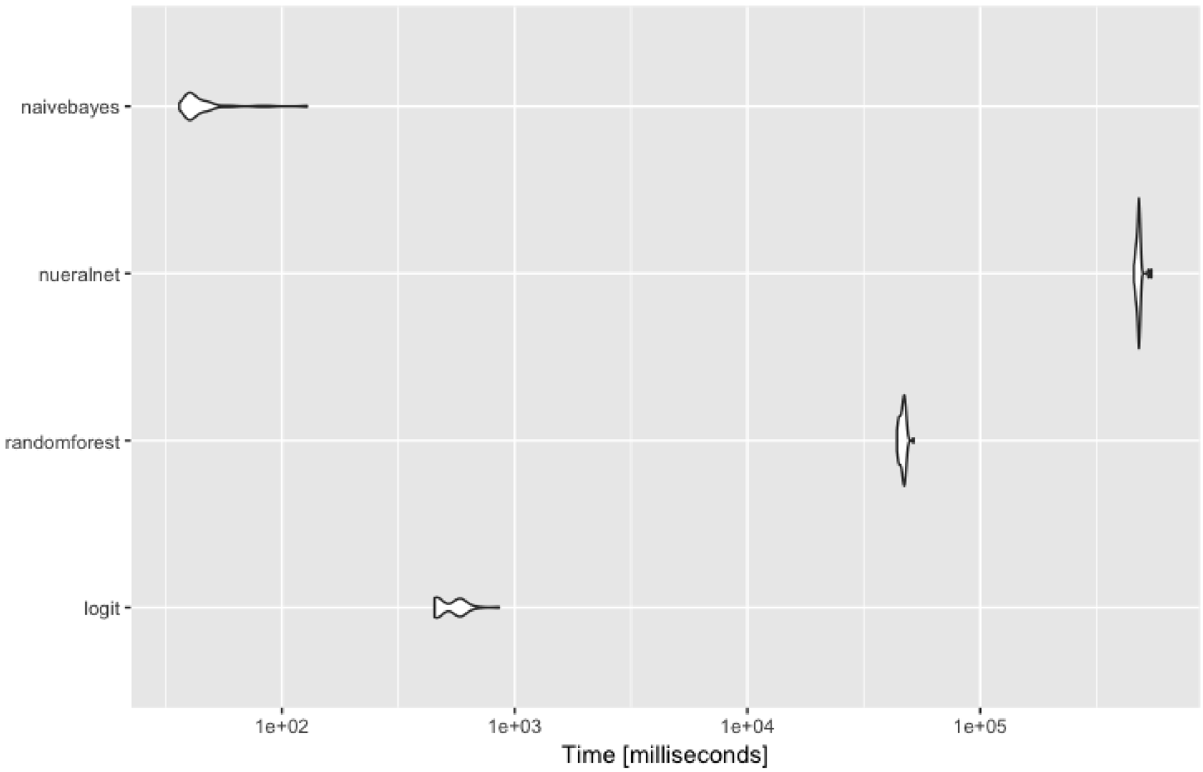

Figure 1: Computational Efficiency